基于并行 C4.5 的铁路零散白货客户流失预测研究 *

张 斌，彭其渊，刘帆洨

(西南交通大学 交通运输与物流学院，成都 610031)

摘 要：为了提高铁路零散白货客户流失预测的准确性和高效性，根据铁路零散白货客户的流失特征，提出了基于 CDL 模型的客户流失识别方法，在此基础上，针对数据量大的问题，提出了基于 Hadoop 并行框架的 C4.5 决策树客户流失预测模型。通过仿真实验，证明该模型具有较好的准确性和预测能力，并且随着样本数量的增加，Hadoop 并行框架的效率得到了明显的提升，且不影响客户流失预测模型的准确性和预测能力。

关键词：铁路运输；零散白货；客户流失；C4.5 决策树；并行；Hadoop

中图分类号：U294.1 doi: 10.3969/j.issn.1001-3695.2017.09.0912

Research on railway scattered freight customer churn prediction based on parallel C4.5 decision tree algorithm

Zhang Bin, Peng Qiyuan, Liu Fanxiao

(*School of Transportation & Logistics*, *Southwest Jiaotong University*, *Chengdu* 610031, *China*)

Abstract: In order to improve the accuracy and efficiency of customer churn prediction of railway scattered freight, according to the loss characteristics of railway scattered freight customers, proposed a customer churn identification method based on CDL model. On this basis, facing the problem of big data, proposed a C4.5 decision tree customer churn prediction model based on Hadoop parallel framework. Simulation results show that the model has good accuracy and predictive ability, and as the number of samples increases, the efficiency of Hadoop parallel framework is obviously improved, and the accuracy and prediction ability of churn prediction model are not affected.

Key Words: railway transportation; scattered freight; customer churn; C4.5 decision tree ; parallel; Hadoop

0 引言

随着全球经济的快速发展，以及国家供给侧改革、“一带一路”发展战略的深化推进和经济结构的有序调整，货物运输市场需求发生了重大变化，逐渐从以大宗货物运输为主向零散白货运输的方向发展，运输组织模式逐渐从以货车编组计划为基础向以客户需求导向为中心的模式发展。然而由于铁路货运在实效性和便捷性等方面存在不足，加上公路、航空等其他运输方式的不断发展壮大，铁路零散白货运输市场面临着激烈的竞争。铁路部门自 2005 年以来，零散白货运输所占份额成逐年下降势态，严重影响了铁路货运市场的地位和收益[1]，据不完全统计，目前国内快递运输 80%采用公路运输，15%采用航空运输，只有 5%采用铁路运输[2]。保证企业核心竞争力的关键是抓住客户[3]，而获取一位新客户的成本是留住一位老客户的 5~6 倍[4-5]。在铁路零散白货运输市场内忧外患的情况下，如何能够最大限度的对货运客户进行管理，从而有效识别可能流失的货运客户，并对其制定挽留策略是铁路货运行业保证核心竞争力的关键，也是提升自身竞争力的有效途径。

目前，在客户流失预测方面的研究方法主要包括统计分析法和人工智能方法[6]，使用最为广泛的算法包括 Logistic 回归[7]、人工神经网络[8]、决策树[9]、支持向量机（SVM）[10]等，其中决策树是通过对训练集进行归纳学习，从无序、无规律的事例中推理生成树状数据结构或决策规则，并运用于新的数据集来进行分类预测的数据挖掘方法。因其较高的准确率以及良好的容脏和解释能力[11]被广泛应用于分类、预测、规则提取等领域。其中 C4.5[12]决策树算法是对 ID3[13]算法的改进算法，其弥补了 ID3 算法中信息增益趋向与多值属性的缺陷。然而决策树是通过迭代计算构成的，当面临大规模数据时，其在计算时间和空间上存在局限，从而严重影响运行效率。Google 公司提供的 Hadoop 分布式开源计算框架能够处理大规模数据，其提供了 MapReduce 编程模型和 Hadoop 分布式文件系统 HDFS，并提供可容错的并行运算方式，可以在其架构上建立大型集群来

基金项目：中国铁路总公司科研计划重大课题（2016X008-J）

作者简介：张斌（1985-），男，内蒙古呼伦贝尔人，博士研究生，主要研究方向为铁路货运大数据分析、计算机仿真，计算机学习（zbin0470@163.com）；彭其渊（1963-），男，教授，博导，主要研究方向为交通运输规划与管理、系统工程、综合运输、物流工程；刘帆洨（1985-），女，重庆长寿人，博士研究生，主要研究方向为交通运输系统工程.

处理大数据集。文献[14]设计并实现了基于 Hadoop 平台的并行 SPRINT 分类算法，并证明其具有较好的分类正确率、较低的时间复杂度和较好的并行性能。文献[15]提出了一种基于 Hadoop 的并行共享决策树挖掘算法，证明其具有良好的并行性和拓展性。文献[16]提出基于 Hadoop 平台的不确定概率 C4.5 算法，并证明其具有处理海量数据的能力。文献[17]提出了基于 Hadoop 的不确定概率误差剪枝算法，并应用于 C4.5 算法中，结合 MapReduce 程序设计，证明其具有处理大规模数据的能力和较好的可扩展性。

本文通过提取零散白货客户货运特征，建立客户的流失识别方法，并针对铁路货运数据量大的问题，采用 C4.5 决策树算法，提出基于 Hadoop 分布式并行架构的零散白货客户流失预测模型，通过仿真实验，证明并行算法的高效性和预测模型的准确性。

1　零散白货客户流失预测模型构建

1.1　零散白货客户流失的识别方法

相对大宗物资，零散白货运价更高，因此零散白货市场是铁路货运市场的高端产品，加上零散白货客户对货运市场服务及动态更加敏感，更加灵活，因此如何提取货运客户流失特征对零散白货客户的流失状态进行判断是对其进行流失预测的重要问题。本文结合零散白货运输特征，从运到期限、货损货差、服务质量三个方面对客户是否具有流失倾向进行判断。运到期限体现了运输时间的兑现率，由于铁路货运过程需要经过发货、途中运输、途中解编、到达四个作业环节，各个环节又有若干操作，往往会影响运到期限，而运到期限是否被满足，对客户是否信赖铁路运输有重要影响。货损货差率是衡量零散白货客户满意度的重要因素，与大宗货物运输不同，零散白货客户对货物的完整性及包装的完好性都提出了较高要求。服务质量表现为客户在铁路货运业务办理流程中的感知和体验，可以从客户的投诉建议情况上得到反馈。

根据零散白货运输特征，本文提出零散白货客户流失识别模型 CDL，其中 C 为在观察窗口内客户的投诉数量，D 为客户单次发货的延误小时数，L 为客户单次发货的货损货差率。V_x 表示在观察窗口内，客户发货延误的平均时间、平均货损货差率、平均客户投诉率。

$$V_x = \frac{\sum_{i=1}^{n} x_i}{F} \quad x \in \{C, D, L\} \tag{1}$$

其中：x_i 表示客户第 i 次发货的客户投诉建议数量 C_i、延误时间 D_i、货损货差率 L_i，F 为观察窗口内客户的发货频率（即发货次数）。

对于客户流失识别 CDL 模型，采用 AHP，并结合德尔菲法对其各项指标赋予权值 $[\omega_C, \omega_D, \omega_L] = [4, 3.3, 2.7]$。从而，得到基于 CDL 模型的零散白货流失因子计算方法，如下所示：

$$G_{CDL}^{j} = \omega_C \times \bar{V}_C^j + \omega_D \times \bar{V}_D^j + \omega_L \times \bar{V}_L^j \tag{2}$$

其中：G_{CDL}^{j} 表示第 j 个客户基于 CDL 模型的流失因子；ω_C、ω_D、ω_L 表示 C、D、L 参数的权值；$\bar{V}_C^j$、$\bar{V}_D^j$、$\bar{V}_L^j$ 分别表示第 j 个客户标准化后的 V_C、V_D、V_L 值。本文采用 Min-max 标准化方法，将各参数的标准化值映射到[0,1]区间，方法如下：

$$\bar{V}_i^j = \frac{V_i^j - \min_{1\le\ell\le F}\{V_i^\ell\}}{\max_{1\le\ell\le F}\{V_i^\ell\} - \min_{1\le\ell\le F}\{V_i^\ell\}}$$

$$i \in \{C, D, L\} \tag{3}$$

其中：$\bar{V}_i^j$ 为第 j 个客户第 i 项参数标准化之后的值。

基于以上分析，本文对零散白货客户流失的识别方法作出以下定义。

定义 1　本文讨论的流失客户指代具有流失倾向的（即将流失的）零散白货客户，对长期未办理业务的客户认定为已经流失，不作为本文的研究范畴。

定义 2　本文根据 CDL 模型的流失因子 G_{CDL}^{j} 和标准化后的模型参数 $\bar{V}_C^j$、$\bar{V}_D^j$、$\bar{V}_L^j$ 来识别零散白货流失客户，识别方法如式（4）所示。

$$\begin{cases} \bar{V}_C^j \ge 0.65 \,||\, \bar{V}_D^j \ge 0.65 \,||\, \bar{V}_L^j \ge 0.75 \\ G_{CDL}^{j} \ge 6.5 \end{cases} \tag{4}$$

如果客户在 CDL 模型中有参数的 $\bar{V}_C^j$、$\bar{V}_D^j$、$\bar{V}_L^j$ 值超过给定阈值的，识别为流失客户；对于未超过阈值的客户，如果流失因子 G_{CDL}^{j} 超过了给定阈值，则识别为流失客户。对于流失的客户标记为 1，未流失的客户标记为 0。

1.2　铁路零散白货客户流失预测模型

在零散白货客户流失识别方法中，本文从运到期限、货损货差、服务质量三个方面对客户进行流失识别判断，但无法预测具有流失倾向的客户。本章结合货运客户的货运特征，在观察窗口内，从客户的注册时长（R）、客户发货频率（F）、客户近期发货表现（N）、客户发货周转量（Z）四个方面，结合并行 C4.5 决策树模型，对零散白货流失客户进行预测研究。其中 $N = {T_{current} - T_{last}}/{T_{average}}$，$T_{current}$ 为观察窗口末端时间，T_{last} 为客户观察窗口内最后一次发货下单时间，$T_{average}$ 为观察窗口内客户的平均发货下单时间间隔；Z 为发货量与发货运距的乘积。

1.2.1 C4.5 决策树

C4.5 决策树的思路是通过计算变量属性的最大信息增益率，来确定决策树从根节点到叶子节点的树状结构，信息增益率最大的变量作为根节点，每一个叶子节点都代表了一类决策结果。

确定决策树的关键是计算各变量属性的最大信息增益率，首先要计算训练样本的信息熵，其表达式如式（5）所示。

$$Info(S) = -\sum_{i=1}^{k} p(i|S)\log_2 p(i|S) \tag{5}$$

其中：S 为训练数据集，$p(i|S)$ 表示 S 中所属类 i 的比例，k

表示类别数量。如果将训练数据集 S 按照属性 A 进行划分，则在已知属性 A 的前提下，S 的不确定度如式（6）所示。

$$Info_A(S) = -\sum_{i=1}^{j}\frac{|S_i|}{|S|}Info(S_i) \tag{6}$$

其中：属性 A 将数据集 S 分为 j 类，$S=\{S_1,S_2,\cdots,S_i,\cdots,S_j\}$，通过计算划分前后的差值，可以得到信息增益，其计算公式为

$$Gain(A,S) = Info(S) - Info_A(S) \tag{7}$$

为了弥补信息增益趋向与多值属性，C4.5 使用信息增益率来克服这个缺陷，如式（8）所示。

$$GainRate(A,S) = \frac{Gain(A,S)}{SplitInfo(A,S)} \tag{8}$$

其中：分割信息量 $SplitInfo(A,S) = -\sum_{i=1}^{j}\left(\frac{|S_i|}{|S|}\right)\log_2\left(\frac{|S_i|}{|S|}\right)$。C4.5 选择具有最大信息增益率的属性，从上往下完成决策树的构建过程。

1.2.2 基于 Hadoop 的并行 C4.5 决策树客户流失预测模型

构建决策树是反复迭代的过程，面对大规模铁路货运零散白货客户信息，如果使用串行计算方式会在运算时间和空间上浪费大量资源。本文基于 Hadoop 分布式平台，使用 MapReduce 计算框架和分布式文件系统 HDFS，建立基于并行 C4.5 决策树的客户流失预测模型。具体操作步骤包括数据源整合加载、数据预处理、流失客户识别、流失客户预测几部分，如图 1 所示。

a)对客户数据源进行整合，包括客户的个人信息、发货信息等，加载进入 HDFS，从而由多数据源转换为单数据源。

b)数据预处理。客户数据信息从 HDFS 中被提取出来，并分割成若干 Split，MapRecuce 使用 JobTracker 将 Split 分配给空闲的 TrackTracker，再由 TrackTracker 分配给 Map 和 Reduce 子任务。Map 接收的数据为<key,value>结构，其执行对客户信息的数据清理操作，包括过滤重复数据、剔除非法数据、过滤无关数据、处理不完整数据及异常数据等。之后将清洗过的数据交由 Reduce 子任务，Reduce 将相同 key 值（客户 ID）的 value 值（客户信息）进行合并，计算客户发货频率 F，并对各项数据进行 Min-max 标准化操作，最后将处理过的数据返回 HDFS。

c)客户流失识别。对客户样本数据进行基于 CDL 模型的客户流失识别。客户样本数据从 HDFS 中提取出来，Map 子任务以<客户 ID,CDL 模型参数>的数据结构发送给 Reduce 子任务，Reduce 对数据以客户 ID 为 key 值进行合并，即以客户为单位汇总观察窗口内的多笔发货信息，计算客户的 $\overline{V}$ 及 G_{CDL} 值，并根据定义二，判断并标识客户流失状态，最后将处理后的客户信息返回 HDFS。

d)流失客户预测。该部分由两个 MapReduce 构成，第一个 MapReduce 过程中，Map1 子任务输入<属性名,（属性值、所属类别、主键 ID）>结构的数据，其中属性名主要由 R、F、N、Z 和流失标识状态组成。Map1 将数据发送给 Reduce1，由于 R、F、N、Z 为连续属性，所以 Reduce1 要对属性值进行 k-means 聚类离散化操作，本文设置 $K=3$，并对所属类别进行计数。第二个 Mapreduce 过程中，将<（属性名，所属类别），（属性值、主键 ID、类别数量）>作为 Map2 的数据读入对象，Reduce2 计算各属性的信息熵和信息增益，并将最大信息增益的属性作为最佳分裂属性，并逐一确定决策树各节点，最终完成决策树的构建。

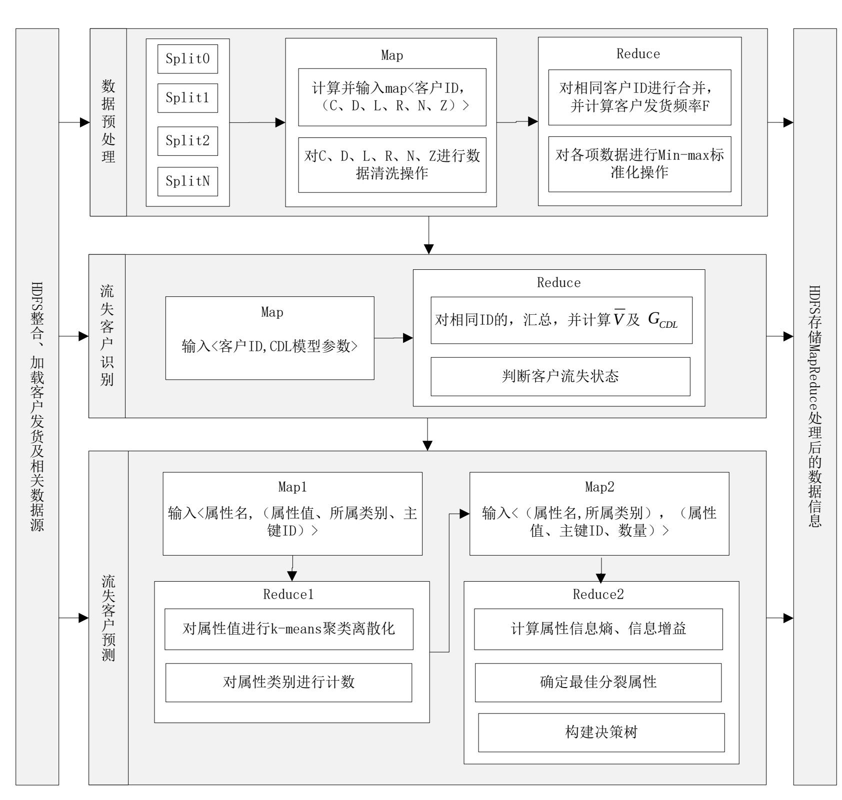


图 1　基于 Hadoop 的并行 C4.5 决策树客户流失预测模型操作步骤

1.2.3 流失预测模型评估标准

混沌矩阵反映了模型的预测效果，是构建模型评估指标的基础[18]。客户流失模型预测结果的混淆矩阵如表 1 所示，其显示了在真实和预测两个维度上客户流失模型预测结果分类。

表 1　客户流失模型预测结果混淆矩阵

样本客户真实状态	预测状态	
	预测流失	预测非流失
真实流失	W	X
真实非流失	Y	U

本文在此基础上，引入模型预测准确率、命中率、覆盖率、提升系数作为评价标准，其定义如式（9）～（12）所示。模型预测准确率为模型整体预测能力；命中率表示正确识别流失客户数占预测为流失客户总数的比例；覆盖率表示正确识别流失客户数占实际流失客户总数的比例；提升系数表示与不利用模型相比，模型预测能力的提升程度。

$$\text{预测准确率}=\frac{W+U}{W+X+Y+U} \tag{9}$$

$$\text{命中率}=\frac{W}{W+Y} \tag{10}$$

$$\text{覆盖率}=\frac{W}{W+X} \tag{11}$$

$$\text{提升系数}=\frac{\text{命中率}}{\text{样本数据中的客户流失率}} \tag{12}$$

2　仿真求解及分析

2.1　仿真数据

本文随机抽取 2016 年全国铁路零散白货货运数据信息、投诉建议系统信息、货运客户数据信息作为仿真数据，每条货运数据为客户单笔发货信息，共计 18 745 208 条，其中包含了客户运到期限完成情况、货损货差情况、客户投诉建议信息、客户单笔周转量、客户发货下单时间、客户基础信息等信息数据。

2.2　模型实现

仿真平台使用局域网，配置 5 台 PC 机作为服务器节点，每台 PC 机装有虚拟机，并搭载 Linux 操作系统，同时配有 4 GB 内存和 500 GB 硬盘存储。每台 PC 机安装了基于 Linux 的 Java 开发包 JDK，并安装了 Hadoop 版本为 2.7.3。仿真平台采用 Hadoop YARN 模式，使用 1 台 PC 为 master，另外 4 台 PC 为 slave 的配置。

模型运行步骤 b)c)后，将货运数据由单笔发货记录汇总为观察窗口内的以客户为单位的客户发货信息，从而得到零散白货客户 27 361 人，其中包括流失客户 8 047 人。C4.5 决策树随机抽取 70%的样本信息作为训练集，另外 30%的样本作为测试集。

2.3　仿真结果及分析

为了验证 C4.5 决策树对铁路零散白货客户流失预测模型具有较高的运算效率和预测效果，以及基于 Hadoop 平台的并行算法的高效性，设计了三个仿真实验来进行验证。

实验 1　在单节点的运行环境下，对比 C4.5 算法、Logistic 算法和 BP 算法的执行效率。在仿真数据中抽取不同数量的样本数据，如表 2 所示。运用三种算法对样本数据进行计算，表 3 显示了三种算法在各样本数据下的运行时间，从中可以看出，三种算法在运行时间上相近，但 C4.5 较其他两种算法在运行速度上略有提升，说明 C4.5 具有较好的运算效率。表 4 显示了运用三种算法运算，结合式（9）（11）得到的客户流失预测模型的准确性和覆盖性，结果显示 C4.5 算法对不同样本的预测在准确性和覆盖性方面都较其他两种算法有优势，说明 C4.5 算法在铁路零散白货客户流失预测模型上具有较好的预测效果。

表 2　样本数据表

样本数据	样本客户数量/个	非流失客户数量/个	流失客户数量/个
D1	800	103	697
D2	2 000	612	1 388
D3	5 000	1 026	3 974
D4	10 000	3 538	6 462

表 3　三种算法运行时间对比结果

	C4.5（s）	Logistic（s）	BP（s）
D1	0.66	0.67	0.73
D2	3.26	3.31	3.50
D3	9.86	9.40	10.11
D4	20.37	20.51	21.01

表 4　三种算法运行时间对比结果/%

	C4.5		Logistic		BP	
	准确率	覆盖率	准确率	覆盖率	准确率	覆盖率
D1	80.63	75.73	79.50	73.79	74.88	60.19
D2	78.30	80.07	77.35	78.76	68.95	65.52
D3	79.80	80.51	79.44	79.82	66.12	60.92
D4	78.87	81.26	77.26	77.47	65.75	56.70

实验 2　在 Hadoop 平台下，对比不同数量服务节点的运行情况。表 5 中显示了在不同数量的服务节点上运行不同数量的仿真样本所需要的运行时间，从表中可以看到，在样本数量较少的情况下，单机模式与并行模式之间的差距很小，然而随着样本数量的增加，基于 Hadoop 的并行运算效率得到了大幅提升，并且服务节点的增加也会随着样本数量的增加大幅降低运算时间。图 2 表示不同节点下，对不同样本进行运算的加速比曲线图，其中，加速比 $\delta=\Omega/\mho$，Ω 表示在单机上运行时间，$\mho$ 表示多节点并行运行时间，其是衡量并行算法的重要参数[19]，

从中可以看出在数据量较少的情况下，加速比变化并不明显，然而随着样本数据的增加，加速比攀升较大，同样说明基于 Hadoop 的并行算法在处理大数据方面具有较大的优势。

表 5　不同服务节点运行不同样本数量的仿真性能

数据量/条	不同服务节点数运行时间/s			
	1 个	2 个	3 个	4 个
10 000	21.37	18.32	17.16	16.44
100 000	128.26	25.38	22.96	19.81
1000 000	987.84	135.76	94.65	56.43
10000 000	内存不足	501.62	297.59	132.97

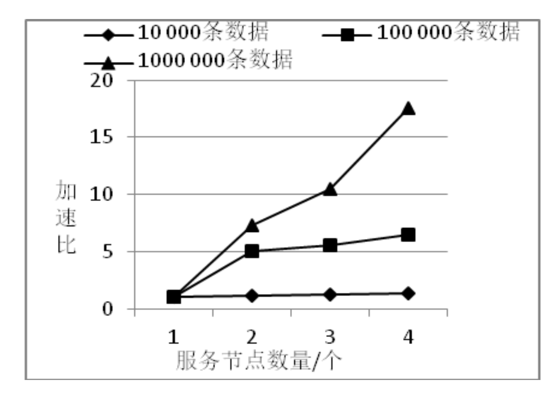


图 2　不同服务节点加速比曲线图

实验 3　在不同数量节点下，运用式（9）~（12）的评估方法，对并行 C4.5 决策树客户流失预测模型进行评估发现，结果如表 6 所示，对于在不同数量的服务节点进行并行实验，模型在准确率、命中率、覆盖率、提升系数等方面都表现良好，说明该模型具有较强的预测能力，并且基于 Hadoop 的并行 C4.5 决策树客户预测模型在不同数量服务节点的情况下，差距不大，说明设定不同节点对于模型的准确性和预测能力影响很小，但是在运行速度方面却有较大提升。

表 6　并行 C4.5 决策树客户流失预测模型评估结果

节点数/个	准确率/%	命中率/%	覆盖率/%	提升系数
1	83.17	69.86	75.20	2.3753
2	84.65	71.32	79.98	2.4250
3	83.13	69.12	77.07	2.3502
4	83.29	70.33	74.70	2.3913

运用并行 C4.5 决策树对仿真数据进行客户流失预测，最终构建的决策树显示，客户的平均发货频率对零散白货客户流失影响最大，为根节点，对于标准化后的平均发货频率小于 0.21 的客户为流失客户，大于 0.73 的客户为非流失客户，其他客户进入决策树第二层分支节点。二层分支节点为客户的平均周转量，对于大于 0.65 的客户为非流失节点，小于 0.13 的客户为流失客户，其他客户的判断进入第三层分支节点，即客户的近期发货表现。对于近期发货表现数据大于 0.82 的客户为流失客户，小于 0.09 的客户为非流失客户，其他客户进入第四层分支节点，即客户的注册时长。对于客户大于 0.75 的客户为非流失客户，其他的为流失客户。通过分析结果可以看出，对客户进行流失预测的因素中，从影响程度上划分，从重到轻依次为客户平均提报频率、平均发货周转量、近期发货表现、注册时长，发货频繁的客户其稳定性越强，而注册时间越久的客户，并不能代表其流失的可能性越小。

3　结束语

本文根据铁路零散白货客户特征，针对零散白货客户流失问题进行了研究，通过建立客户流失识别 CDL 模型，并对客户流失因子进行计算，定义了零散白货客户的流失识别方法，之后运用大数据技术，建立了基于 Hadoop 分布式平台和 C4.5 决策树的客户流失预测模型，并使用 MapReduce 计算框架和分布式文件系统 HDFS 对模型进行了仿真求解，结果显示 C4.5 算法对铁路零散白货客户流失预测模型的计算较 Logistic 算法和 BP 算法具有较高的运算效率和精确度，并且基于 Hadoop 并行计算方法使得算法运算效率得到了大幅提升，并且客户流失预测模型的准确性和预测能力没有受到影响，对于大数据量的测试样本具有较大的实用价值。该方法可以有效指导铁路货运部门对零散白货客户流失进行预测，从而有针对性的制定客户挽留策略，实现铁路货运增运增收的目的。

参考文献：

[1] 王志美, 张星臣, 徐彬. 零散白货的货源组织问题和运输组织问题一体化 [J]. 北京交通大学学报, 2016, 40 (6): 43-49+56.

[2] 张伯敏. 供给侧改革下铁路从大宗货物向零散快捷货物拓展的思考 [J]. 交通运输工程与信息学报, 2016, 14 (4): 36-40.

[3] 周新军. 客户关系管理引入铁路货运服务的理论与实践 [J]. 铁道货运, 2008, 26 (12): 25-28.

[4] Athanassopoulos A D. Customer satisfaction cues to support market segmentation and explain switching behavior [J]. Journal of Business Research, 2000, 47 (3): 191-207

[5] Bhattacharya C B. When customers are members: customer retention in paid membership contexts [J]. Journal of the Academy of Marketing Science, 1998, 26 (1): 31-44.

[6] 夏国恩, 金炜东. 基于支持向量机的客户流失预测模型 [J]. 系统工程理论与实践, 2008, 28 (1): 71-77.

[7] Chang Chengchang, Gong Dahchuan. A comparison of rohs risk assessment using the logistic regression model and artificial neural network model [C]// Proc of the 9th International Conference on Machine Learning and Cybernetics. 2010.

[8] 余路. 电信客户流失的组合预测模型 [J]. 华侨大学学报: 自然科学版, 2016, 37 (5): 637-640.

[9] 叶志龙, 黄章树. 线上会员客户流失的建模与预测研究 [J]. 管理现代化, 2016, 36 (3): 96-98.

[10] 于小兵, 卢逸群. 电子商务客户流失预警与预测 [J]. 系统工程, 2016

(9): 37-43.

[11] 张宇, 张之明. 一种基于 C5.0 决策树的客户流失预测模型研究 [J]. 统计与信息论坛, 2015, 30 (1): 89-94.

[12] Quinlan J R. C4. 5: programs for machine learning [M]. San Francisco: Morgan Kaufmann Publishers, 1993: 17-42.

[13] QUINLAN J R. Induction of decision trees [J]. Machine Learning, 1986, 1 (1): 81-106.

[14] 黄刚, 孙媛. 基于 Hadoop 平台的 SPRINT 算法的分析与研究 [J]. 南京师大学报: 自然科学版, 2016, 39 (4): 25-30.

[15] 陈湘涛, 张超, 韩茜. 基于 Hadoop 的并行共享决策树挖掘算法研究 [J]. 计算机科学, 2013, 40 (11): 215-221.

[16] 刘亚秋, 李海涛, 景维鹏. 基于 Hadoop 的海量嘈杂数据决策树算法的实现 [J]. 计算机应用, 2015, 35 (4): 1143-1147.

[17] 张晶星, 李石君. 基于 Hadoop 的改进决策树剪枝算法 [J]. 计算机工程与设计, 2016, 37 (7): 1942-1946.

[18] 贺本岚. 支持向量机模型在银行客户流失预测中的应用研究 [J]. 金融论坛, 2014, 225 (9): 70-74.

[19] 陆秋, 程小辉. 基于 MapReduce 的决策树算法并行化 [J]. 计算机应用, 2012, 32 (9): 2463-2465, 2469.